

DATA SET MODELS AND EXPONENTIAL FAMILIES IN STATISTICAL PHYSICS AND BEYOND

Jan Naudts and Ben Anthonis

University of Antwerp, Physics Department

Universiteitsplein 1, 2610 Wilrijk-Antwerpen, Belgium

jan.naudts@ua.ac.be, ben.anthonis@ua.ac.be

Abstract

The exponential family of models is defined in a general setting, not relying on probability theory. Some results of information geometry are shown to remain valid. Exponential families both of classical and of quantum mechanical statistical physics fit into the new formalism. Other less obvious applications are predicted. For instance, quantum states can be modeled as points in a classical phase space and the resulting model belongs to the exponential family.

1 Introduction

The exponential family of statistical models is an important notion in statistics. The parametrized statistical model $\theta \in \mathbb{R}^n \rightarrow p_\theta(a)$ belongs to the *exponential family*[1] if there exist functions $\alpha(\theta)$, $c(a)$, and $H_j(a)$, $j = 1, 2, \dots, n$, such that the probability distributions $p_\theta(a)$ can be written as

$$p_\theta(a) = c(a) \exp \left[-\alpha(\theta) - \sum_{j=1}^n \theta_j H_j(a) \right]. \quad (1)$$

The choice of signs conforms with the conventions of statistical physics where the Boltzmann-Gibbs probability distribution is usually written as

$$p_\beta(a) = \frac{c(a)}{Z(\beta)} e^{-\beta H(a)}. \quad (2)$$

This distribution is parametrized by the inverse temperature β and clearly belongs to the exponential family. The function $H(a)$ is called the Hamiltonian, the normalization $Z(\beta)$ is called the partition sum. The function $c(a)$ is a prior weight. In many cases it is identically equal to 1. But for instance, if the underlying measure space A is the set of integers \mathbb{N} , then $c(a) = 1/a!$ might be an appropriate choice.

Recently, generalizations of the notion of an exponential family have been introduced[2, 3, 4, 5, 6, 7, 8, 9, 10]. They provide a solid theoretical underpinning for research in non-extensive statistical physics[11, 12]. The equilibrium probability distributions (pdfs) studied in this context are related to Amari's α -family of pdfs[13]. The latter is the subject of research in information geometry[14], where techniques from differential geometry are applied to probability theory.

The present work has been inspired by the efforts of Topsøe[15, 16] to formulate the notion of an exponential family in an abstract setting of game theory. One of his goals is to formulate information theory without involving statistics. From [15] we quote: "In 1983 Kolmogorov stated that 'Information theory must precede probability theory and not be based on it'." A seminal paper in this direction is the work of Csiszár[17]. The settings of this paper can be reformulated in the terminology used in the present work. More recent contributions in the area of machine learning are found in [18, 19].

The next Section introduces the abstract settings of the formalism. In Section 3 the notion of Entropy is added. Section 4 gives a definition of an exponential family of models. Section 5 shows that both the standard and the quantum mechanical notions of an exponential family fit into the present formalism. The final Section formulates some conclusions.

2 Data set models

2.1 The information framework

The elements of our framework are

The *space of data sets* \mathbb{X} is an abstract topological space. Following Topsøe [15, 16] an element x of \mathbb{X} can be called a *truth*. However, it is closer to the tradition of probability theory to consider the space of possible outcomes of an experiment. Therefore we refer to x as a *data set*. In the probabilistic formulation of information theory \mathbb{X} is the space of probability distributions over a finite alphabet A . In the quantum mechanical context it is the space of quantum states, for instance described by normalized wave functions or by density operators. Other examples are given in what follows.

The space of *questions* \mathbb{Q} is a dual space of \mathbb{X} . Each question q is a real function continuously defined on an open subset of \mathbb{X} . The evaluation of q in the point x is the *answer to the question* and is denoted $\langle x|q \rangle$ instead of $q(x)$ to stress that the space of questions is a linear space but not necessarily an algebra with the usual pointwise product. For instance, each hermitian bounded operator A on the Hilbert space of wavefunctions ψ determines an everywhere defined continuous function, given by

$$\psi \rightarrow \langle \psi|A \rangle \equiv (\psi, A\psi). \quad (3)$$

Here, (ϕ, ψ) is the scalar product of two elements ϕ, ψ of the Hilbert space $\mathcal{L}^2(\mathbb{R}^3, \mathbb{C})$. Note that we follow the notational conventions of the physics literature. In the case of an unbounded operator, such as the position operators or many of the Hamilton operators, some caution is needed. One must select a topology which makes (3) continuous on the domain of definition of the operator.

2.2 What is a model?

In statistical physics a model is determined by its Hamiltonian. In the present context this is replaced by one or more questions. However, we want to make the definition slightly more general by introducing the following definition.

Definition 1 A data set model is a topological manifold¹ \mathbb{M} together with a continuous map μ defined on an open subset of the space \mathbb{X} of data sets taking values in \mathbb{M} .

Clearly, a set of questions q_1, \dots, q_n with a common open domain of definition D defines a manifold $\mathbb{M} \subset \mathbb{R}^n$ as the range of the map μ defined by $\mu(x) = U$ when $U_j = \langle x|q_j \rangle, j = 1, 2, \dots, n$, provided that the set $\mu(D)$ is open in \mathbb{R}^n .

The converse is also true. Indeed, one has

Proposition 1 A local parametrization $U \in D \subset \mathbb{R}^n \rightarrow m_U \in \mathbb{M}$ of the manifold \mathbb{M} , μ defines questions q_j by $\langle x|q_j \rangle = U_j$ when $\mu(x) = m_U$.

Proof

The questions are well-defined. The domain of definition is the set of x for which $\mu(x)$ belongs to the range of the map $U \in D \rightarrow m_U \in \mathbb{M}$. This is an open set because any homeomorphism is an open map. It is also bijective so that there is a unique U such that $m_U = \mu(x)$. Hence, the answer to the questions q_j is unique.

The map $x \rightarrow \langle x|q_j \rangle = U_j$ is continuous because $\mu(x)$ is continuous and $U \in D \rightarrow m_U \in \mathbb{M}$ is open. □

The advantage of defining a model in terms of manifolds is that the dependence on a specific choice of questions has been eliminated.

Example 1 The Euclidean space $\mathbb{X} = \mathbb{R}^3$ is a space of data sets. The unit sphere

$$S_2 = \{x \in \mathbb{R}^3 : |x| = 1\} \quad (4)$$

is a model embedded in \mathbb{R}^3 . The map μ is defined on $\mathbb{R}^3 \setminus \{0\}$ by $\mu(x) = x/|x|$. The questions q_1 and q_2 defined for $x_3 > 0$ by

$$\langle x|q_1 \rangle = \frac{x_1}{x_3} \quad \text{and} \quad \langle x|q_2 \rangle = \frac{x_2}{x_3}. \quad (5)$$

determine a parametrization of the northern hemisphere of S_2 . It is given by

$$U \rightarrow x_U = (U_1 x_3, U_2 x_3, x_3)^T \quad \text{with} \quad x_3 = \frac{1}{\sqrt{1 + U_1^2 + U_2^2}}. \quad (6)$$

3 Maximum entropy principle

3.1 Entropy functions

The amount of information contained in the data set x is given by its *entropy* $S(x)$. It is a lower semi-continuous function² with values in the extended reals $[-\infty, +\infty]$. Usually the entropy is assumed to be concave. However, in general the space \mathbb{X} does not have an affine structure. On

¹ \mathbb{M} is locally Euclidean, this means that there exists in each point m of \mathbb{M} an integer $n > 0$, an open set D of \mathbb{R}^n , together with a map $U \in D \rightarrow x_U \in \mathbb{M}$ which is a homeomorphism between D and a neighbourhood of m .

²We do not use this property in the present paper.

the other hand, models are manifolds. Hence, by transferring the notion of entropy to the model points the concavity as a function of parameters can be discussed.

Given a data set model \mathbb{M}, μ the entropy $S(m)$ of a model point m is defined by the *maximum entropy principle* of Jaynes[21]

$$S(m) = \sup\{S(x) : \mu(x) = m\} \leq +\infty. \quad (7)$$

If m is not in the range of μ then $S(m) = -\infty$ is chosen. Note that we use here the map μ as a constraint on the data sets involved in the maximization procedure, instead of using a specific set of questions q_1, \dots, q_n .

Since \mathbb{M} is a manifold we can now investigate whether local parametrizations $U \rightarrow m_U$ exist such that $S(m_U)$ is a concave function of the parameters U . In what follows the notation $S(U) \equiv S(m_U)$ will be used. Note that $S(U)$ depends on the choice of local parametrization while $S(m)$ is independent of parametrization.

Proposition 2 *Let $U \in D \subset \mathbb{R}^n \rightarrow m_U$ be a local parametrization of a data set model \mathbb{M}, μ , Let q_1, \dots, q_n be the accompanying set of questions as defined by Proposition 1. Then one has locally*

$$S(U) = \sup\{S(x) : \langle x|q_j \rangle = U_j \text{ for } j = 1, 2, \dots, n\} \leq +\infty. \quad (8)$$

The proof of this result is straightforward.

Example 2 *Consider the parametrization of the northern hemisphere of the unit circle, as discussed before. The entropy function*

$$S(x) = -1 - |x|(\ln |x| - 1) \quad (9)$$

is maximal when $|x| = 1$. The entropy function $S(m)$ vanishes on the model manifold.

3.2 Perfect data sets

In the example of the sphere the supremum in (7) is actually a maximum. The entropy function $S(x)$ takes on its maximal value for the points of S_2 . It is then obvious to call these points *perfect data sets*. Such privileged data points do not always exist. For instance, the model for a quantum particle can be a point particle localized at a position q in \mathbb{R}^3 . The map μ is defined by $\mu(\psi) = \langle \psi|Q\psi \rangle$. But there are no quadratically integrable wavefunctions which describe a quantum particle perfectly localized at the position q . In such a case one expects an entropy function $S(\psi)$ which is such that no maximum is attained for any wave function ψ .

The relation between model points and perfect data sets may be a one-to-many relation. This is made clear in the following example.

Example 3 *In the case of linear regression a data set consists of a finite sequence of pairs of real numbers*

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n), \quad (10)$$

with at least two distinct pairs. The model space consists of straight lines not parallel to the y -axis. A data set is perfect if the data points fall on a single line. But with a single straight line correspond many perfect data sets. See the Figure 1.

The interesting questions are given by

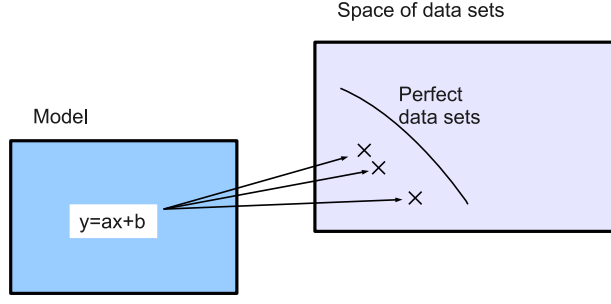


Figure 1: Embedding of the model into the space of data sets.

- $q_a(x, y) = \frac{1}{Z} \sum_{i,j} (y_i - y_j)(x_i - x_j);$
- $q_b(x, y) = \frac{1}{Z} \sum_{i,j} \sum_{i,j} (x_i y_j - x_j y_i)(x_i - x_j).$

with $Z = \sum_{i,j} (x_i - x_j)^2$. They are only defined on data sets for which $Z \neq 0$. They are interesting because they return the parameters a and b of the fitted line $y = ax + b$. These two questions uniquely determine the model. A meaningful entropy function is

$$S(x, y) = -\frac{1}{Z} \sum_{i,j=1}^n (x_i y_j - x_j y_i)^2 - \frac{1}{Z} \sum_{i,j}^n (y_i - y_j)^2. \quad (11)$$

Its value on perfect data sets is $-a^2 - b^2$. For other data sets is $S(x) < S(\mu(x))$.

4 Exponential families

The notion of an exponential family of models is strongly related to the concept of canonical parametrizations. These are introduced now.

4.1 Contact transforms

In thermodynamics, the Massieu function $\Phi(\theta)$ is the Legendre transform of the entropy $S(U)$. This inspires for the following definition.

Definition 2 Let be given a local parametrization $U \in D \subset \mathbb{R}^n \rightarrow m_U$ of a data set model \mathbb{M}, μ . Assume that the model entropy $S(U)$ is locally finite. Then the Massieu function is defined by

$$\Phi(\theta) = \sup_{U \in D} \left\{ S(U) - \sum_{j=1}^n \theta_j U_j \right\}. \quad (12)$$

Theorem 1 Let be given a local parametrization $U \in D \subset \mathbb{R}^n \rightarrow m_U$ of a data set model \mathbb{M}, μ . Let q_1, \dots, q_n be the accompanying set of questions defined by Proposition 1. Assume that the model entropy $S(U)$ is locally finite. Then one has

$$\Phi(\theta) = \sup \{ S(x) - \sum_{j=1}^n \theta_j \langle x | q_j \rangle : \mu(x) \text{ is local} \}. \quad (13)$$

$\Phi(\theta)$ is a convex function. In particular, it is finite on a convex subset Θ of \mathbb{R}^n .

Proof

Remember that the questions are such that $\mu(x) = m_U$ holds if and only if $\langle x|q_j \rangle = U_j$ for $j = 1, 2, \dots, n$. Take x so that $\mu(x)$ is local. Then one has $\mu(x) = m_U$ with $U \in D$. But $S(U) < +\infty$ implies that $S(x) < +\infty$. Hence one has

$$S(x) - \sum_{j=1}^n \theta_j \langle x|q_j \rangle \leq S(U) - \sum_{j=1}^n \theta_j U_j \leq \Phi(\theta). \quad (14)$$

On the other hand, if $\Phi(\theta) < +\infty$ then for any $\epsilon > 0$ there exists U such that

$$\Phi(\theta) - \epsilon < S(U) - \sum_{j=1}^n \theta_j U_j. \quad (15)$$

Similarly, there exists x , satisfying $\langle x|q_j \rangle = U_j$ for $j = 1, 2, \dots, n$, such that

$$S(U) - \epsilon < S(x). \quad (16)$$

All together one has

$$\Phi(\theta) - 2\epsilon < S(x) - \sum_{j=1}^n \theta_j U_j. \quad (17)$$

Since $\epsilon > 0$ is arbitrary one concludes that the equality holds in (13).

Finally, if $\Phi(\theta) = +\infty$ then there exists U such that $S(U) - \sum_{j=1}^n \theta_j U_j$ is arbitrary large. But then there exists x such that $\mu(x)$ is local and $S(x) - \sum_{j=1}^n \theta_j \langle x|q_j \rangle$ is arbitrary large. Hence, also in this case the equality holds in (13).

The convexity statement is easy to show. Let λ in $[0, 1]$. One can assume that $\Phi(\theta_1)$ and $\Phi(\theta_2)$ are finite because otherwise the convexity statement is empty. Then for any x with local $\mu(x)$ one has

$$\begin{aligned} & S(x) - \sum_{j=1}^n [\lambda \theta_{1,j} + (1 - \lambda) \theta_{2,j}] \langle x|q_j \rangle \\ &= \lambda \left[S(x) - \sum_{j=1}^n \theta_{1,j} \langle x|q_j \rangle \right] + (1 - \lambda) \left[S(x) - \sum_{j=1}^n \theta_{2,j} \langle x|q_j \rangle \right] \\ &\leq \lambda \Phi(\theta_1) + (1 - \lambda) \Phi(\theta_2). \end{aligned} \quad (18)$$

This implies $\Phi(\lambda \theta_1 + (1 - \lambda) \theta_2) \leq \lambda \Phi(\theta_1) + (1 - \lambda) \Phi(\theta_2)$. □

In the physics literature one is used to work with the free energy rather than with Massieu's function. If the inverse temperature β is the only parameter then the free energy equals $-\Phi(\beta)/\beta$ and minimizes $\langle x|q \rangle - S(x)/\beta$.

4.2 Canonical parametrization

Let us now return to a data set model with a locally defined parametrization. Then the Legendre-Fenchel transform can be used to introduce a canonical parametrization. The attribute 'canonical' refers to the canonical ensemble of statistical physics. In the context of the exponential family one speaks about the canonical form of the probability distribution. But in the present approach the canonical parametrization is defined before introducing the exponential family and is independent of it.

Definition 3 *Let be given some local parametrization $\theta \in \Theta \subset \mathbb{R}^n \rightarrow m_\theta$ of a data set model \mathbb{M}, μ . The parametrization is said to be canonical if there exists another local parametrization $U \in D \subset \mathbb{R}^n \rightarrow m_U$ such that*

- $S(U) < +\infty$ for all U in D ;
- The relation $m_\theta = m_U$ defines a diffeomorphism between Θ and D ;
- Under this diffeomorphism is

$$\Phi(\theta) - S(U) + \sum_{j=1}^n \theta_j U_j = 0. \quad (19)$$

To make the distinction between the two parametrizations $\theta \in \Theta \subset \mathbb{R}^n \rightarrow m_\theta$ and $U \in D \subset \mathbb{R}^n \rightarrow m_U$ we call the latter the associated energy parametrization. The motivation is that in statistical physics the components of U have the meaning of energies.

Theorem 2 *If the parametrization $\theta \in \Theta \rightarrow m_\theta$ of a data set model \mathbb{M}, μ is canonical then the Massieu function $\Phi(\theta)$ is a strictly convex differentiable function and there exist questions q_1, \dots, q_n satisfying*

$$\frac{\partial}{\partial \theta_j} \Phi(\theta) = -\langle x | q_j \rangle \quad \text{for all } x \text{ satisfying } \mu(x) = m_\theta. \quad (20)$$

Proof

Let $U \in D \subset \mathbb{R}^n \rightarrow m_U$ be the local parametrization appearing in the definition of a canonical parametrization. Note that

$$\zeta \rightarrow \Phi(\theta) - \sum_{j=1}^n U_j (\zeta_j - \theta_j) \quad (21)$$

is a tangent plane in the point θ . The requirement that $m_U = m_\theta$ determines a diffeomorphism implies that a small change of θ corresponds with a small change of U and hence a small change in the slope of the tangent plane. This proves that the tangent plane is unique. One concludes that $\Phi(\theta)$ is differentiable and that

$$\frac{\partial \Phi}{\partial \theta_j} = -U_j. \quad (22)$$

The strict convexity follows because the correspondence $\theta \leftrightarrow U$ is bijective.

Let q_1, \dots, q_n be the questions defined in Proposition 1. They satisfy $\langle x|q_j \rangle = U_j$ for $j = 1, 2, \dots, n$ when $\mu(x) = m_U$. Hence the statement of the Theorem follows. \square

The second derivatives of $\Phi(\theta)$ define a metric tensor

$$g_{j,k}(\theta) = \frac{\partial^2 \Phi}{\partial \theta_j \partial \theta_k} = -\frac{\partial U_k}{\partial \theta_j}. \quad (23)$$

This matrix is a generalization of Fisher's information matrix.

Example 4 Let \mathbb{X} be the set of all 2-by-2 density operators (these are positive trace class operators with trace equal to 1). The entropy function is the von Neumann entropy

$$S(\rho) = -\text{Tr } \rho \ln \rho. \quad (24)$$

The model \mathbb{M} coincides with the space of data sets \mathbb{X} . Let us calculate a parametrization which is canonical.

Three questions are needed to determine uniquely a density operator ρ . In terms of the three Pauli matrices σ_j these are

$$\langle \rho|q_j \rangle = \text{Tr } \rho \sigma_j, \quad j = 1, 2, 3. \quad (25)$$

Then one can write

$$\rho = \frac{1}{2} \left(\mathbb{I} + \sum_j U_j \sigma_j \right) \quad \text{with } U_j = \langle \rho|q_j \rangle. \quad (26)$$

The von Neumann entropy becomes

$$S(\rho) = \ln 2 - \frac{1}{2}(1 + |U|) \ln(1 + |U|) - \frac{1}{2}(1 - |U|) \ln(1 - |U|). \quad (27)$$

The Massieu function reads

$$\Phi(\theta) = \sup_U \left\{ S(U) - \sum_{j=1}^3 \theta_j U_j : |U| \leq 1 \right\}. \quad (28)$$

The maximum is reached when

$$\theta_j = \frac{1}{2} \frac{U_j}{|U|} \ln \frac{1 - |U|}{1 + |U|}. \quad (29)$$

Note that this implies that $|U| = \tanh |\theta|$. Hence the inverse relation is

$$U_j = -\frac{\theta_j}{|\theta|} \tanh |\theta|. \quad (30)$$

One concludes that the map $U \rightarrow \theta$ is a diffeomorphism from the interior of the unit sphere onto \mathbb{R}^3 .

ρ_θ can now be written as

$$\begin{aligned} \rho_\theta &= \frac{1}{2} \mathbb{I} - \frac{1}{2|\theta|} \tanh |\theta| \sum_{j=1}^3 \theta_j \sigma_j \\ &= \frac{1}{2 \cosh(|\theta|)} \exp \left(- \sum_j \theta_j \sigma_j \right). \end{aligned} \quad (31)$$

This is a canonical parametrization of the 2-by-2 density matrices.

4.3 Dual Relations

Let be given a canonical parametrization $\theta \rightarrow m_\theta$ of model \mathbb{M}, μ , together with the associated energy parametrization $U \rightarrow m_U$. From (19, 20) then follows the pair of dual relations

$$\frac{\partial \Phi}{\partial \theta_j} = -U_j \quad \text{and} \quad \frac{\partial S}{\partial U_j} = \theta_j, \quad (32)$$

where $U \rightarrow \theta$ is the diffeomorphism determined by the relation $m_U = m_\theta$.

The function $S(U)$ is strictly concave. This follows because the matrix of second derivatives of $S(U)$ equals minus the inverse of the metric tensor $g_{j,k}(\theta)$ defined by (23). The latter is positive definite because by Theorem 2 the Massieu function is strictly convex.

If the metric tensor $g_{j,k}(\theta)$ is sufficiently smooth then the model space \mathbb{M} is (locally) a Riemannian manifold with respect to each of the two parametrizations. They are dual to each other in the sense that the metric tensor of one parametrization is the inverse of that of the other. The curvature of the manifold in the Levi-Civita connection vanishes because the metric tensor is the matrix of second derivatives of a convex function. Hence the manifold is flat.

4.4 Logarithmic maps

Definition 4 *A logarithmic map L maps model points onto questions.*

For instance, the Boltzmann-Gibbs-Shannon entropy $S(p)$ can be written as the average of the measurable quantity $-\ln p(i)$. The probability distribution p belongs to the space of data sets \mathbb{X} . But $-\ln p(i)$ is used as a question, the answer of which is the value of the entropy function $S(p)$. In this example the logarithmic map is defined on all data sets. But we need it further on only for perfect data sets or for model points.

The logarithmic map L can be used to define a *divergence* or *relative entropy* between data sets and model points.

Definition 5 *The divergence of a data set x from a model point m is given by*

$$D(x||m) = \sup\{S(y) + \langle y|Lm \rangle : \mu(y) = m\} - S(x) - \langle x|Lm \rangle. \quad (33)$$

Clearly, if $\mu(x) = m$ then $D(x||m) \geq 0$ with equality if and only if x maximizes $S(x) + \langle x|Lm \rangle$ under the constraint $\mu(x) = m$. We call such x *canonical* data sets.

4.5 Exponential families

In the previous subsection the notion of a logarithmic map was introduced to prepare for the definition of the exponential family.

Definition 6 *A model \mathbb{M}, μ with logarithmic map L belongs to the exponential family of data set models if the model space \mathbb{M} is covered with local parametrizations $\theta \in \Theta \rightarrow m_\theta$, which are canonical, and the associated energy parametrizations $U \in D \subset \mathbb{R}^n \rightarrow m_U$ are such that*

$$Lm_\theta = \alpha(\theta) - \sum_j \theta_j q_j \quad \text{for all } \theta \in \Theta, \quad (34)$$

where the questions q_j are defined by $\langle x|q_j \rangle = U_j$ when $\mu(x) = m_U$ (see Proposition 1).

In the example of the 2-by-2 density matrices (see (31)) is

$$\ln \rho_\theta = -\ln 2 \cosh(|\theta|) - \sum_j \theta_j \sigma_j. \quad (35)$$

Hence the model belongs to the exponential family. One has $\alpha(\theta) = -\ln 2 \cosh(|\theta|)$. The questions q_j are given by (25).

The property (34) can be used to simplify the Definition 5 of divergence. One obtains

$$\begin{aligned} D(x||m_\theta) &= \sup\{S(y) + \langle y|\alpha(\theta) - \sum_j \theta_j q_j \rangle : \mu(y) = m_\theta\} \\ &\quad - S(x) - \langle x|\alpha(\theta) - \sum_j \theta_j q_j \rangle \\ &= \sup\{S(y) - \langle y|\sum_j \theta_j q_j \rangle : \mu(y) = m_\theta\} \\ &\quad - S(x) + \sum_j \theta_j \langle x|q_j \rangle \\ &= \Phi(\theta) - S(x) + \sum_j \theta_j \langle x|q_j \rangle. \end{aligned} \quad (36)$$

From Theorem 1 now follows that $D(x||m_\theta) \geq 0$ for all x for which $\mu(x)$ is local. Equality then holds if and only if the data set is canonical.

Note that one can write, using (19),

$$D(x||m_\theta) = S(U) - \sum_j \theta_j U_j - \left[S(x) - \sum_j \theta_j \langle x|q_j \rangle \right]. \quad (37)$$

If $\mu(x) = m_U$ then $\langle x|q_j \rangle = U_j$. Hence

$$D(x||m_\theta) = S(U) - S(x) \geq 0 \quad \text{if } \mu(x) = m_U. \quad (38)$$

Therefore, in the case of a model belonging to the exponential family, canonical data sets are perfect data sets as well.

4.6 Pythagorean Theorems

The model map μ can be seen as an orthogonal projection of \mathbb{X} onto the manifold \mathbb{M} . This is supported by a Pythagorean theorem in which the divergence plays the role of a distance squared.

Introduce the divergence between two model points m and m' by

$$D(m||m') = \inf\{D(x||m') : \mu(x) = m\}. \quad (39)$$

The following result shows that this divergence is of the Bregman type[17, 20]. It has a nice geometric interpretation. It is the difference between the value $\Phi(\zeta)$ of the Massieu function in the point ζ and the value of the plane tangent in the point θ .

Proposition 3 *Let be given a model \mathbb{M}, μ with logarithmic map L belonging to the exponential family. Consider a local parametrization $\theta \in \Theta \rightarrow m_\theta$ and the associated energy parametrization $U \in D \subset \mathbb{R}^n \rightarrow m_U$ as in the definition of the exponential family. Then one has*

$$D(m_\theta||m_\zeta) = \Phi(\zeta) - \Phi(\theta) + \sum_j (\zeta_j - \theta_j) U_j. \quad (40)$$

Proof

First calculate using (36)

$$\begin{aligned} D(m_\theta||m_\zeta) &= \inf\{D(y||m_\zeta) : \mu(y) = m_\theta\} \\ &= \Phi(\zeta) - \sup\{S(y) - \sum_j \zeta_j \langle y|q_j \rangle : \mu(y) = m_\theta\}. \end{aligned} \quad (41)$$

Now use that $\langle y|q_j \rangle$ is constant on the set of y for which $\mu(y) = m_\theta$. Hence one has

$$D(m_\theta||m_\zeta) = \Phi(\zeta) - S(U) + \sum_j \zeta_j U_j \quad (42)$$

with U so that $m_U = m_\theta$. Using (19) this becomes (40). □

The Pythagorean theorem[17] for the projection of an arbitrary data set $x \in \mathbb{X}$ onto the manifold \mathbb{M} by means of the model map μ now follows readily. See the Figure 2.

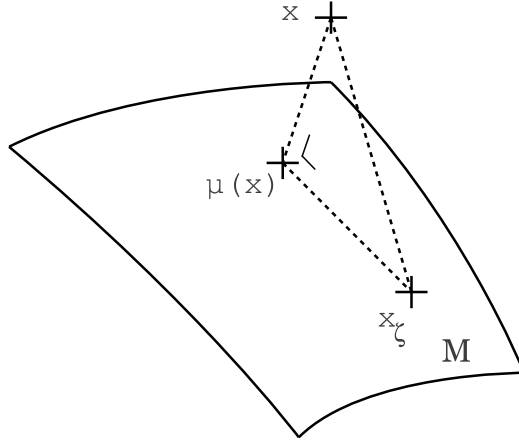


Figure 2: Projection of a data set x onto the manifold \mathbb{M} using the model map μ .

Theorem 3 *Let be given a model \mathbb{M}, μ with logarithmic map L belonging to the exponential family. If $\mu(x) = m_\theta$ then*

$$D(x||m_\theta) + D(m_\theta||m_\zeta) = D(x||m_\zeta). \quad (43)$$

Proof

Use (40) to obtain

$$D(x||m_\theta) + D(m_\theta||m_\zeta) = \Phi(\zeta) - S(x) + \sum_j \zeta_j \langle x|q_j \rangle = D(x||m_\zeta). \quad (44)$$

This is (43). □

Following [10], we can also formulate a Pythagorean theorem involving only model points.

Theorem 4 Consider a model \mathbb{M}, μ with logarithmic map L belonging to the exponential family. Let $\theta \in \Theta \rightarrow m_\theta$ and $U \in D \rightarrow m_U$ be canonical and energy parametrizations as mentioned in Definition 6. Let θ, ζ, ξ be points in Θ . Let U and V be dual coordinates such that $m_U = m_\theta$ and $m_V = m_\zeta$. Assume that

$$\sum_j (\zeta_j - \xi_j)(U_j - V_j) = 0. \quad (45)$$

Then one has

$$D(m_\theta || m_\zeta) + D(m_\zeta || m_\xi) = D(m_\theta || m_\xi). \quad (46)$$

Proof

This follows immediately from (40). □

5 Applications

We show below how the standard notion of an exponential family of statistical models fits into the present formalism. Also the analogue notion in quantum statistics is discussed. The generalized exponential families[2] introduced in the context of Tsallis' non-extensive statistical mechanics[12], or even in a broader context, do fit as well, but will not be treated here.

5.1 Statistical models

Here we show that the above framework is a generalization of the notion of the exponential family of statistical models[1].

Let \mathbb{X} be the affine space of probability distributions over the discrete measure space A . Let $c(a)$ be a prior weight on A . Questions are real functions f of A , seen as maps $p \rightarrow \mathbb{E}_p f = \sum_a p(a)f(a)$. The answer to a question f , given p , is therefore given by

$$\langle p | f \rangle = \mathbb{E}_p f. \quad (47)$$

The entropy function is that of Boltzmann-Gibbs-Shannon (BGS) and is given by

$$S(p) = -\langle p | L(p) \rangle = -\sum_a p(a) \ln \frac{p(a)}{c(a)}. \quad (48)$$

Let $\theta \in \Theta \rightarrow p_\theta$ be a statistical model with probability distributions p_θ given by (1). For convenience assume $c(a) = 1$ and introduce the notation $\mathbb{E}_\theta \equiv \mathbb{E}_{p_\theta}$. Let $U_j(\theta) = \mathbb{E}_\theta H_j$. The model space \mathbb{M} is the subset of \mathbb{X} given by

$$\mathbb{M} = \{p_\theta : \theta \in \Theta\}. \quad (49)$$

Introduce the model map μ by

$$\mu(p) = p_\theta \quad \text{if } \mathbb{E}_p H_j = \mathbb{E}_\theta H_j \text{ for } j = 1, \dots, n. \quad (50)$$

Assume for convenience that the functions $H_j(a)$ are bounded. Then the model map is everywhere defined and continuous in the l_1 -metric of \mathbb{X} .

It is well-known that the probability distributions of a model belonging to the exponential family maximise the BGS-entropy under the constraint $U_j(\theta) = \mathbb{E}_\theta H_j$, $j = 1, \dots, n$ — in our terminology the p_θ are perfect data sets. Hence one has

$$S(\theta) = S(U) = S(p_\theta) = \alpha(\theta) + \sum_j \theta_j U_j(\theta). \quad (51)$$

In particular, there follows that $\Phi(\theta) = \alpha(\theta)$.

Generically, the relation between U and θ is a diffeomorphism. Indeed, one has

$$\begin{aligned} g_{j,k}(\theta) &= -\frac{\partial U_k}{\partial \theta_j} \\ &= -\frac{\partial}{\partial \theta_j} \sum_a p_\theta(a) H_k(a) \\ &= \sum_a p_\theta(a) H_j(a) H_k(a) + \sum_a p_\theta(a) \frac{\partial \alpha}{\partial \theta_j} H_j \\ &= \mathbb{E}_\theta H_j H_k - (\mathbb{E}_\theta H_j) (\mathbb{E}_\theta H_k). \end{aligned} \quad (52)$$

If the constant function is not a linear combination of the hamiltonians H_j then the matrix $g_{j,k}(\theta)$ is positive definite. This implies that the relation between U and θ is a diffeomorphism.

One concludes that the parametrization $\theta \rightarrow p_\theta$ is canonical.

Introduce a logarithmic map L by

$$(Lp_\theta)(a) = \ln p_\theta(a). \quad (53)$$

The corresponding divergence is

$$D(p||p_\theta) = \sum_a p(a) \ln \frac{p_\theta(a)}{p(a)}. \quad (54)$$

This is the standard expression for the divergence/relative entropy.

It follows now from (51) that the model \mathbb{M}, μ with this logarithmic map belongs to the exponential family provided that no linear combination of the hamiltonians H_j is a constant function.

5.2 Quantum statistical physics

In quantum statistics the probability distributions of classical statistics are replaced by density matrices/density operators on a separable Hilbert space. They form the space \mathbb{X} of data sets. Questions are bounded operators on the Hilbert space. The evaluation function is

$$\rho \in \mathbb{X} \rightarrow \langle \rho | A \rangle \equiv \text{Tr } \rho A. \quad (55)$$

It is continuous for instance in the Hilbert-Schmidt norm. The entropy function is the von Neumann entropy (24).

A quantum statistical model is a homeomorphism $\theta \in \Theta \subset \mathbb{R}^n \rightarrow \rho_\theta$. The model space is $\mathbb{M} = \{\rho_\theta : \theta \in \Theta\}$. The model belongs to the exponential family of quantum models if there exist self-adjoint operators H_1, \dots, H_n such that

$$\rho_\theta = \frac{1}{Z(\theta)} \exp\left(-\sum_{j=1}^n \theta_j H_j\right) \quad (56)$$

with $Z(\theta) = \text{Tr} \exp(-\sum_{j=1}^n \theta_j H_j)$. The model map μ satisfies $\mu(\rho) = \rho_\theta$ if $\text{Tr} \rho H_j$ is well-defined and equals $U_j = \text{Tr} \rho_\theta H_j$ for $j = 1, \dots, n$.

The ρ_θ of the form (56) maximize the von Neumann entropy under the constraint of a given value of the U_j . The proof is based on Klein's inequality — see for instance [22, 9]. In particular the ρ_θ are perfect data sets. One obtains

$$S(U) = S(\rho_\theta) = \Phi(\theta) + \sum_{j=1}^n \theta_j U_j \quad \text{with} \quad \Phi(\theta) = \ln Z(\theta). \quad (57)$$

One calculates

$$\begin{aligned} g_{j,k}(\theta) &= -\frac{\partial U_k}{\partial \theta_j} = -\frac{\partial}{\partial \theta_j} \text{Tr} \rho H_k \\ &= \text{Tr} \rho H_j H_k - \frac{\partial Z}{\partial \theta_j} \text{Tr} \rho H_k \\ &= \text{Tr} \rho H_j H_k - (\text{Tr} \rho H_j)(\text{Tr} \rho H_k). \end{aligned} \quad (58)$$

The eigenvalues of this matrix cannot be negative. If they are strictly positive for all θ then the relation between U and θ is a diffeomorphism and the parametrization $\theta \rightarrow \rho_\theta$ is canonical.

Introduce the logarithmic map defined by $L\rho_\theta = \ln \rho_\theta$. One clearly has

$$L\rho_\theta = -\ln Z(\theta) - \sum_{j=1}^n \theta_j H_j. \quad (59)$$

Hence, the model belongs to the exponential family according to Definition 6. A short calculation then yields

$$D(\rho || \rho_\theta) = \text{Tr} \rho (\ln \rho - \ln \rho_\theta). \quad (60)$$

This is the standard expression for relative entropy in quantum statistical physics[23].

5.3 Coherent states

Now we discuss an example which shows that our framework extends well beyond the (quantum) statistical context. We consider the phase space of classical mechanics as a model for a state space of quantum mechanical wave functions.

For simplicity consider a quantum particle in one dimension. The space \mathbb{X} of data sets consists of wave functions $\psi(x)$ which are twice differentiable and normalized so that

$$\int_{\mathbb{R}} dx |\psi(x)|^2 = 1. \quad (61)$$

Note that two wave functions $\psi(x)$ and $e^{i\alpha\psi(x)}$, with α constant, determine the same point of \mathbb{X} .

Questions are linear operators A acting on the Hilbert space of square integrable complex functions. The evaluation function is given by

$$\langle \psi | A \rangle = \int_{\mathbb{R}} dx \overline{\psi(x)} (A\psi)(x). \quad (62)$$

Introduce position and momentum operators by $Q\psi(x) = x\psi(x)$ and $P\psi(x) = -i\hbar\frac{\partial\psi}{\partial x}$. Note that these are unbounded operators. Hence we need a topology on \mathbb{X} which is such that the two questions $\psi \rightarrow \langle\psi|Q\rangle$ and $\psi \rightarrow \langle\psi|P\rangle$ are continuous. Then they define a continuous map μ of \mathbb{X} into the model space $\mathbb{M} = \mathbb{R}^2$, which is the phase space of a particle in classical mechanics.

Introduce now the entropy function

$$S(\psi) = \frac{1}{2} |\langle\psi|a\rangle|^2 - \langle\psi|a^\dagger a\rangle, \quad (63)$$

where the annihilation operator a is defined by

$$a = \frac{1}{\sqrt{2}} \left[\frac{1}{r}Q + i\frac{r}{\hbar}P \right], \quad (64)$$

with r and \hbar positive constants. Then \mathbb{X} together with this entropy function is a data set space.

The solution of the eigen equation $a\psi = z\psi$, with complex z , is denoted ψ_z and is called a *coherent state*. Note that

$$U_1 = \langle\psi_z|Q\rangle = r\sqrt{2}\Re z \quad \text{and} \quad U_2 = \langle\psi_z|P\rangle = \frac{\hbar}{r}\sqrt{2}\Im z. \quad (65)$$

and

$$|\langle\psi|a\rangle|^2 = \frac{1}{2r^2}U_1^2 + \frac{r^2}{2\hbar^2}U_2^2. \quad (66)$$

Clearly is

$$S(\psi_z) = -\frac{1}{2} |\langle\psi_z|a\rangle|^2 = -\frac{1}{2}|z|^2, \quad (67)$$

and

$$S(\psi) \leq -\frac{1}{2} |\langle\psi|a\rangle|^2 \quad \text{for all } \psi \in \mathbb{X} \text{ for which } \langle\psi|a\rangle = z. \quad (68)$$

Hence, the coherent states are perfect data sets. In particular, the entropy $S(m)$ of the model point $m = m_U$ is

$$S(U) = -\frac{1}{2r^2}U_1^2 - \frac{r^2}{2\hbar^2}U_2^2. \quad (69)$$

The Massieu function equals

$$\Phi(\theta) = \sup_U \{S(U) - \theta_1 U_1 - \theta_2 U_2\}. \quad (70)$$

The maximum is reached when

$$\theta_1 = -\frac{1}{r^2}U_1 \quad \text{and} \quad \theta_2 = -\frac{r^2}{\hbar^2}U_2. \quad (71)$$

The result is

$$\Phi(\theta) = \frac{r^2}{2}\theta_1^2 + \frac{\hbar^2}{2r^2}\theta_2^2. \quad (72)$$

It is now straightforward to verify that the θ -parametrization of \mathbb{R}^2 is canonical.

Introduce a logarithmic map L by

$$L(m_U) = -\frac{1}{2}|z|^2 + \frac{1}{2}za^\dagger + \frac{1}{2}\bar{z}a, \quad (73)$$

where z is obtained from (65). There follows immediately that

$$L(m_U) = -\Phi(\theta) - \theta_1 Q - \theta_2 P. \quad (74)$$

This shows that the model belongs to the exponential family. The divergence equals

$$D(\phi||m_U) = \frac{1}{2}|\langle\phi|a\rangle - z|^2 + \langle\phi|a^\dagger a\rangle - |\langle\phi|a\rangle|^2 \geq 0. \quad (75)$$

In addition, $D(\phi||\psi_z) = 0$ is equivalent with $z = \langle\phi|a\rangle$ and $a\phi = \langle\psi|a\rangle\phi$. But this implies that ϕ equals ψ_z , up to a phase factor which can be neglected because it has no physical meaning. Hence, the divergence vanishes if and only if ϕ equals ψ_z up to a constant phase factor.

6 Conclusions

The notion of an exponential family of models can be generalized to a context not involving probability theory. From the point of view of statistical physics this is of interest because the exponential family is at the heart of the discipline and quantum statistical physics involves quantum probability rather than classical probability theory. But the formalism presented here is so general that it has many other applications. Only one such example has been elaborated in subsection 5.3. Some other applications have been mentioned without proof. These will be taken up in further work.

By the present effort we hope to contribute to a more general theory of information, including previous extensions in the directions of machine learning, statistical inference and quantum information.

References

- [1] O. E. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory* (J. Wiley and Sons, New York, 1978).
- [2] J. Naudts, *Estimators, escort probabilities, and phi-exponential families in statistical physics*, J. Ineq. Pure Appl. Math. **5** (2004) 102.
- [3] P. D. Grünwald and A. P. Dawid, *Game Theory, Maximum Entropy, Minimum Discrepancy And Robust Bayesian Decision Theory*, Ann. Stat. **32** (2004) 1367–1433.
- [4] J. Naudts, *Generalised exponential families and associated entropy functions*, Entropy **10** (2008) 131–149.
- [5] J. Naudts, *The q-exponential family in statistical physics*, Cent. Eur. J. Phys. **7** (2009) 405–413.
- [6] A. Ohara, *Geometric study for the Legendre duality of generalized entropies and its application to the porous medium equation*, Eur. Phys. J. **B70** (2009) 15–28.
- [7] A. Ohara and T. Wada, *Information geometry of q-Gaussian densities and behaviors of solutions to related diffusion equations*, J. Phys. **A43** (2010) 035002.

- [8] J. Naudts, *The q -exponential family in statistical physics*, Proceedings of Kyoto RIMS workshop: "Mathematical Aspects of Generalized Entropies and their Applications", ed. H. Suyari, A. Ohara, T. Wada, J. Phys.: Conf. Series **201** (2010) 012003.
- [9] J. Naudts, *Generalised Thermostatistics* (Springer Verlag, 2011).
- [10] S. Amari and A. Ohara, *Geometry of q -Exponential Family of Probability Distributions*, Entropy **13** (2011) 1170–1185.
- [11] C. Tsallis, *Possible Generalization of Boltzmann-Gibbs Statistics*, J. Stat. Phys. **52** (1988) 479–487.
- [12] C. Tsallis, *Introduction to nonextensive statistical mechanics* (Springer Verlag, 2009).
- [13] S. Amari, *Differential-geometrical methods in statistics*, Lecture Notes in Statistics **28** (1985).
- [14] S. Amari and H. Nagaoka, *Methods of Information Geometry*, Translations of Mathematical Monographs (Oxford University Press, Oxford, UK, 2000).
- [15] F. Topsøe, *Game Theoretical Optimization inspired by Information Theory*, J. Global Optim. **43** (2009) 553–564.
- [16] F. Topsøe, *Elements of the cognitive universe*, <http://www.math.ku.dk/~topsoe/isit2011.pdf> (2011).
- [17] I. Csiszár, *Why least squares and maximal entropy? An axiomatic approach to inference for linear inverse problems*, Ann. Stat. **19** (1991) 2032–2066.
- [18] T. D. Sears, *Generalized Maximum Entropy, Convexity, and Machine Learning*. PhD thesis, Australian National University (2008).
- [19] Nan Ding and S. V. N. Vishwanathan, *t -Logistic regression*, Adv. Neural Inf. Proc. Systems (2010) <http://books.nips.cc/nips23.html>.
- [20] L.M. Bregman, *The relaxation method to find the common point of convex sets and its applications to the solution of problems in convex programming*, USSR Computational Mathematics and Mathematical Physics **7** (1967) 200–217.
- [21] E. Jaynes, *Information theory and statistical mechanics*, Phys. Rev. **106** (1957) 620–630.
- [22] D. Ruelle, *Statistical mechanics, Rigorous results*. (W.A. Benjamin, Inc., New York, 1969).
- [23] D. Petz, *Bregman divergence as relative operator entropy*, Acta Math. Hungar. **116** (2007) 127–131.